AD721700

SDC

TM-4652/300/00

DESCRIPTION AND ANALYSIS OF THE VICENS-REDDY RECOGNITION ALGORITHMS

29 March 1971

D D C
RECEIVED
APR 12 1971
C

Reproduced by
NATIONAL TECHNICAL
INFORMATION SERVICE
Springfield, Va. 22151

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

35

BEST
AVAILABLE COPY

SERIES BASE NO VOL REISSUE

TM 4652 /300/00

TECHNICAL MEMORANDUM

(TM Series)

The work reported herein was supported by the Advanced Research Projects Agency of the Department of Defense under Contract DAHC15-67-C-0149, ARPA Order No. 1327, Amendment No. 3, Program Code No. 1D30, and 1P10.

DESCRIPTION AND ANALYSIS OF THE VICENS-REDDY RECOGNITION ALGORITHMS

by

Iris Kameny

H. Barry Ritea

29 March 1971

SYSTEM DEVELOPMENT CORPORATION
2500 COLORADO AVE.
SANTA MONICA
CALIFORNIA
90406

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the U. S. Government.

Distribution of this document is unlimited.

A-1153 (5/68)

ABSTRACT

This document provides a detailed description and analysis of the recognition algorithms used in the Vicens-Reddy speech recognition system.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

This document provides a detailed description of the recognition procedures used in the Vicens-Reddy speech recognition system [1]. It is a sequel to SDC TM-4652/200, Description and Analysis of the Vicens-Reddy Preprocessing and Segmentetion Algorithms, to which the reader is referred for a description of the terms and variables used.

Recognition is a method of assigning linguistic labels to the sustained and transitional segments of the P-matrix. There are 14 such labels for the sustained segments and one label for the transitional segments. These are given in Table 1.

Table 1. Labels for Transitional and Sustained Segments

Linguistic Label	Four-Character Name	Type Number
Transitional	TRAN	0
Consonant	CNST	1
Nasal	NASL	2
Stop	STOP	3
Burst	BRST	4
Fricative	FRIC	5
Vowel type 1	VWL1	6
Vowel type 2	VWL2	7
Vowel type 3	VWL3	8
Vowel type 4	VWL4	9
Vowel type 5	VWL5	10
Vowel type 6	VWL6	11
Vowel type 7	VWL7	12
Vowel type 8	VWL8	13
Vowel type 9	VWL9	14

Note that most of the conventional linguistic groups of phonemes are included in the table. Other groups, such as glides, have been omitted.*

Recognition is divided into three parts: (1) Primary Classification, (2) Secondary Classification, and (3) Construction of the R-matrix. Primary classification is a serial process in which each sustained segment is first tested to see if it is a fricative; if it is not classified as a fricative, tests are sequentially performed for the following groups:

- Vowel
- Stop
- Nasal
- Consonant

The label "consonant" is attached to all those sustained segments not falling into the other categories. Because of this serial process, the phoneme groups given in Table 1 are not mutually exclusive. For, if a sustained segment satisfies the test for a vowel but could also fulfill the test for a nasal, it would never be considered a nasal since the vowel test precedes that for nasals. Secondary classification regroups adjacent fricatives and adjacent stops and detects and labels burst segments. Special tests are then performed to define beginning and ending segments. Finally, an array called the R-matrix, or feature matrix, is constructed.

2. PRIMARY CLASSIFICATION

Primary classification consists of five steps of sequentially determining (1) fricatives, (2) vowels, (3) stops, (4) nasals, and (5) consonants.

---

*It is felt that if the original transitional segments occurring in secondary segmentation were retained, rather than being extended onto surrounding sustained segments, they might provide a clue for the existence of glides.

2.1 FRICATIVE DETERMINATION

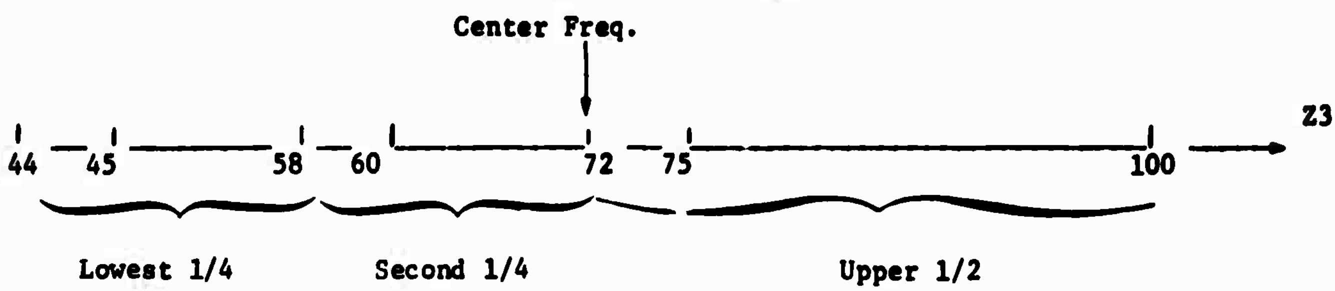
P(i) is labeled a fricative (i.e., TYPE(i) = 5) if either:

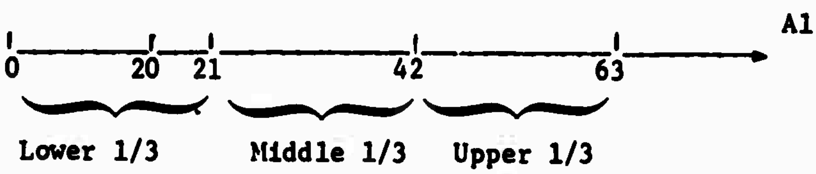
(1) $Z3(i) \geq 75$
and $A1(i) \leq 20$,

or (2) $60 \leq Z3(i) < 75$,
$A3(i) \geq A1(i)$,
and $A1(i) \leq 20$,

or (3) $45 \leq Z3(i) < 60$,
$A1(i) \leq 12$,
and $A3(i) \geq A1(i)$.

In an attempt to explain the above three tests, we note that fricatives are generally characterized by a high Z3 frequency and a low A1 amplitude (e.g., see [3] and [4]). Consider now the following diagram of the Z3 and A1 ranges:





For test (1), Z3(i) must be in approximately the upper half of the third frequency band, and A1(i) must be in the lower third of all possible A1 amplitudes.

For test (2), Z3(i) is required only to be in the second fourth of all possible Z3 values. A1(i) must also be in the lower third of its range as before. However, because the constraint on Z3 has been lowered, an additional condition, that A3(i) ≥ A1(i), has been added.

In test (3), Z3(i) has the nominal requirement to be in the lowest fourth of all Z3 frequencies. However, the test for A1(i) is now made more stringent: A1(i) is now required to be in the lowest 20% of all of its possible values. In addition, we retain the requirement that A3(i) ≥ A1(i) as in test (2).

The condition that A3(i) ≥ A1(i) is illustrated by the energy spectra as given by Heinz and Stevens [4] (see Figure 1). These spectra indicate that the above tests are reasonable for the determination of the fricative /ʃ/. However, they seem inappropriate for a characterization of /s/ since the cutoff for Z3 is 5000 Hz, whereas the spectra indicate that Z3 is actually around 5500 - 8000 Hz.

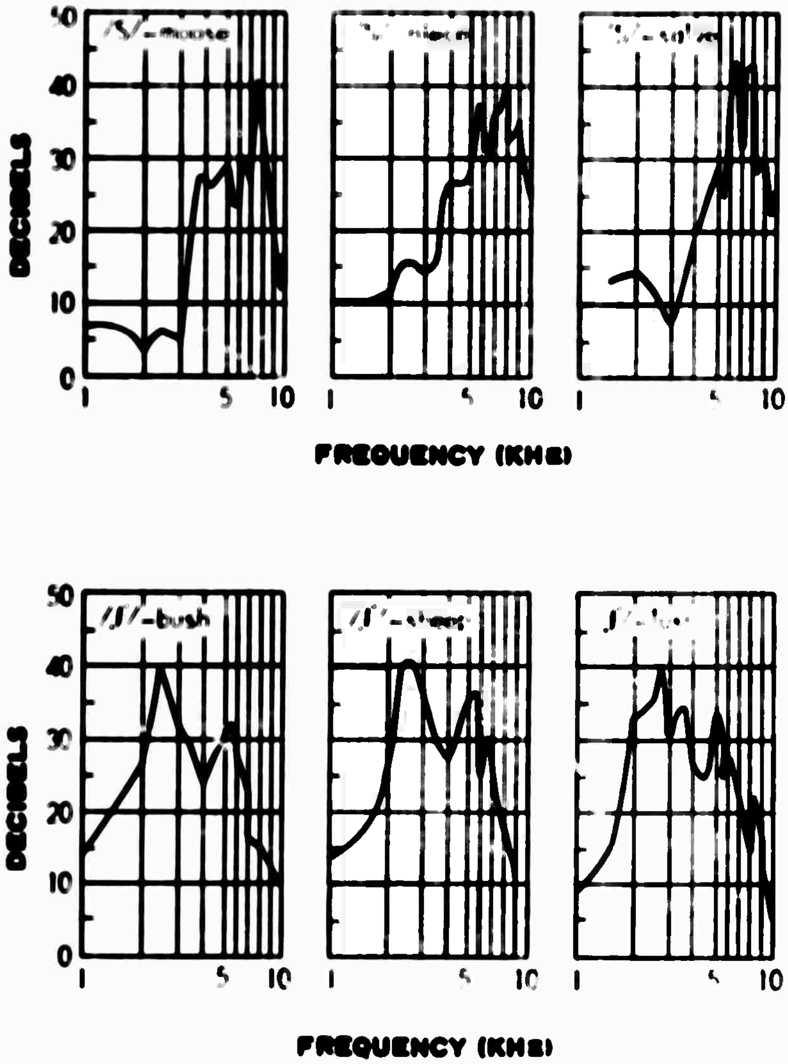


Figure 1. Energy Spectra of the Fricatives /s/ and /ʃ/
(adapted from Heinz and Stevens [4]).

2.2 VOWEL DETERMINATION

If P(i) has not been labeled a fricative, it is labeled a vowel* if:

(1) it is a local maximum (i.e., SXT(i) = 1)

(2) $A1(i) \geq 1$[illegible]

(3) $A1(i) + A2(i) + A3(i) \geq 25$

and

(4) $DUR(i) > 8$.

The test for a vowel as given in the program also requires that

$$5 \cdot DUR(i) + A1(i) + A2(i) + A3(i) \geq 50.$$

However, this condition is superfluous since it is automatically implied by conditions (3) and (4).

Generally, a vowel is characterized in the literature as a speech segment of sufficient duration and amplitude. In the present case, this is characterized by conditions (2), (3), and (4). However, an additional constraint, viz, condition (1), is imposed.

Each P(i) found to be a vowel is assigned a type number TYPE(i) as follows:

$$TYPE(i) = \begin{cases} 6 \text{ if } Z1(i) < 6 \text{ and } Z2(i) < 18 \\ 7 \text{ if } Z1(i) < 6 \text{ and } 18 \leq Z2(i) < 27 \\ 8 \text{ if } Z1(i) < 6 \text{ and } Z2(i) \geq 27 \\ 9 \text{ if } 6 \leq Z1(i) < 9 \text{ and } Z2(i) < 18 \\ 10 \text{ if } 6 \leq Z1(i) < 9 \text{ and } 18 \leq Z2(i) < 27 \\ 11 \text{ if } 6 \leq Z1(i) < 9 \text{ and } Z2(i) \geq 27 \\ 12 \text{ if } Z1(i) > 9 \text{ and } Z2(i) < 18 \\ 13 \text{ if } Z1(i) > 9 \text{ and } 18 \leq Z2(i) < 27 \\ 14 \text{ if } Z1(i) > 9 \text{ and } Z2(i) \geq 27 \end{cases}$$

---

*In searching for a vowel every SXT(i) = 1 is reset to SXT(i) = 0. That is, there are no indicators of local maximums left from this point on. For a detailed description of the meaning of a local maximum, see [2] pp. 18-22.

This is illustrated in Figure 2. If 5 is subtracted from the type number so that the range is changed from 6-14 to 1-9, the type corresponds to the vowel subclasses.

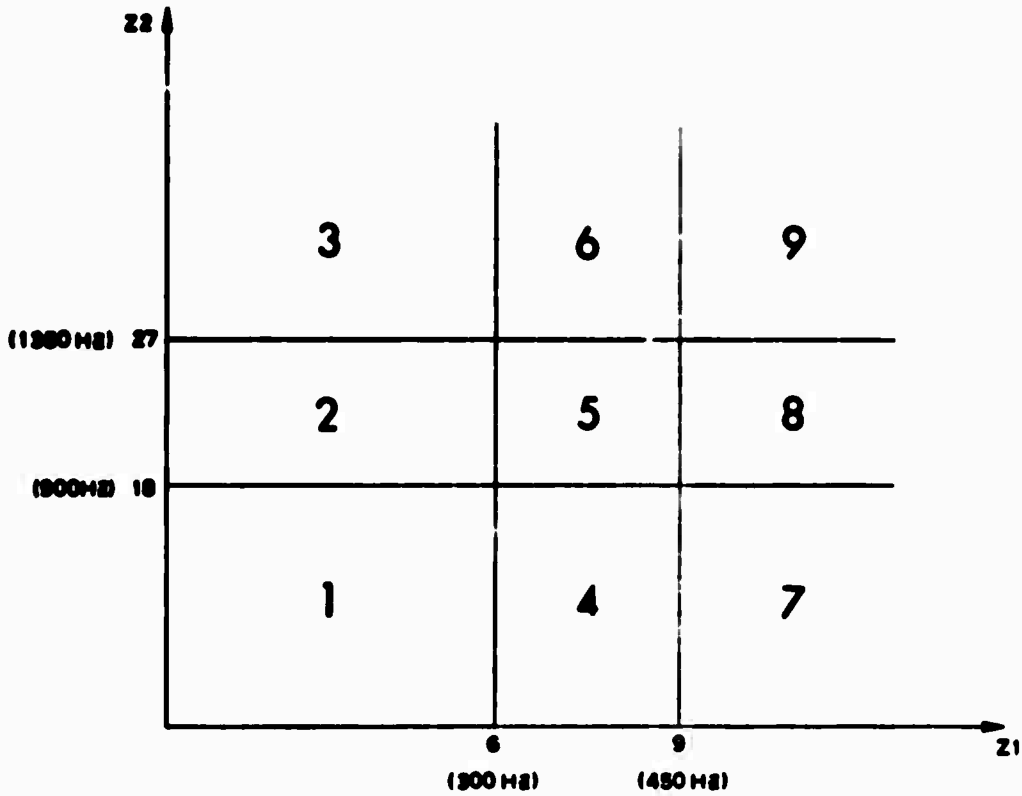


Figure 2. Vowel Subclasses (adapted from Vicens [1])

If tests (1), (2), and (3) are satisfied but (4) is not, so that DUR(i) ≤ 8, then we search the surrounding segments to find the one most likely to be a vowel. To perform this search, we begin by defining

$$\text{AMPLIM} = \text{A1}(i) + \text{A2}(i) + \text{A3}(i) - \frac{\text{A1}(i)+\text{A2}(i)+\text{A3}(i)+4}{3 \cdot \text{DUR}(i)}$$

Then for $j = i-1, i-2, \ldots$, we search <u>backwards</u> from $P(i)$ until a $P(j)$ is found for which either

$$A1(j) < 16$$

or

$$A1(j) + A2(j) + A3(j) < \max\{25, AMPLIM\}$$

or

$$TYPE(j) = FRIC$$

or

$SXT(j) = -1$ (i.e., $P(j)$ is a local minimum).

We then let

$K1 = j+1$ and $DUR1 = DUR(j+1)$.

A <u>forward</u> search is now made for $k = j+1, j+2, \ldots$, SIZEP to find a $P(k)$ for which

$$A1(k) \geq 16$$

and $A1(k) + A2(k) + A3(k) > \max\{25, AMPLIM\}$

and $TYPE(k) \neq FRICS$

and $SXT(k) \neq -1$ (not a local minimum).

Then if $|DUR(k) - DUR1| \leq 2$

and $A1(k) + A2(k) + A3(k) \geq A1(K1) + A2(K1) + A3(K1)$,

or if $|DUR(k) - DUR1| > 2$

and $DUR(k) > DUR1$

then we set $DUR1 = DUR(k)$

and $K1 = k$ and continue our forward search.

But whenever we find a P(k) for which

$A1(k) < 16$

or $A1(k) + A2(k) + A3(k) \leq \max\{25, AMPLIM\}$

or TYPE (k) = FRICS

or SXT(k) = -1 (local minimum)

or k = SIZEP +1,

then we consider P(K1) to be the best choice, and if

$5 \cdot DUR(K1) + A1(K1) + A2(K1) + A3(K1) \geq 50,$

then we let i = K1 and label P(i) a vowel, using the numbers TYPE(i) as given above.

The literature on acoustic phonetics abounds with papers on vowel characterizations. Results from a few representative papers have been selected to help explain Vicens' vowel subclasses. In particular, it is interesting to compare the present vowel classifications with those obtained by Peterson and Barney [5] and Forgie and Forgie [6] (see Figures 3 and 4). A glossary of the phonemic symbols used in Figures 3 and 4 is given in the appendix. A comparison of Figure 2 with Figures 3 and 4 indicates that the vowel classifications used by Vicens do not correlate well with those obtained by either Peterson and Barney or Forgie and Forgie. First of all, Figure 2 indicates nine vowel categories, whereas Figures 3 and 4 show ten. Also, Vicens does not correlate his vowel categories with particular vowel phonemes.

One reason for the poor correlation is due to hardware anomalies in the Vicens-Reddy system. Indeed, zero-crossings are not counted if below the threshold of .03V. This causes the Z1 and Z2 frequencies to be lower than their actual values. These lower frequencies are reflected in the different cut-off values for the vowel categories. In addition, the three fixed front-end filters make it difficult to obtain formant 1 and formant 2 frequencies; i.e., Z1 and Z2 can be poor approximations to the actual formant 1 and formant 2 frequencies.

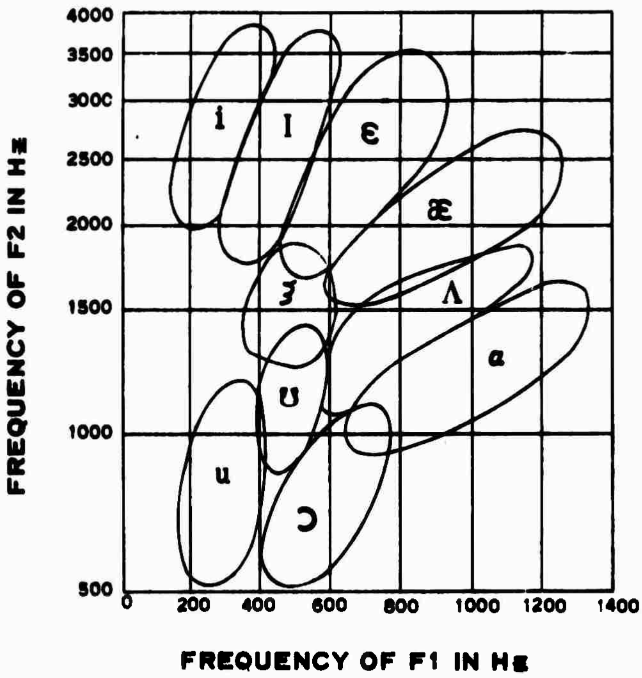


Figure 3. Formant 2 vs. Formant 1 Vowel Plot
(adapted from Peterson and Barney [5])

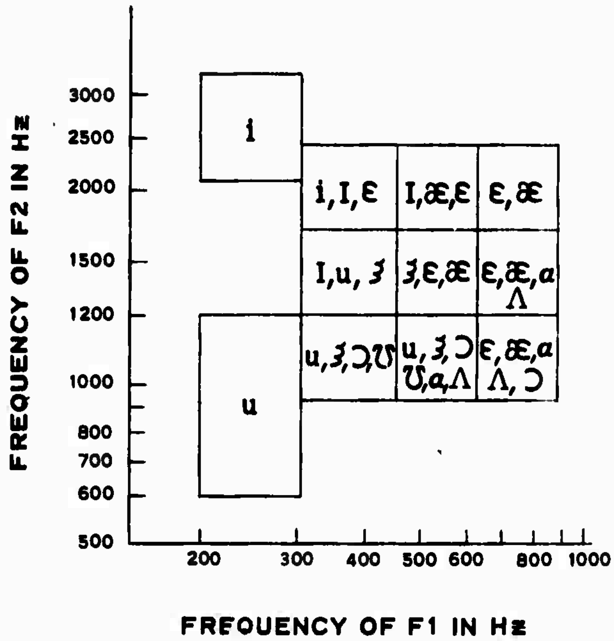


Figure 4. Formant 2 vs. Formant 1 Vowel Plot
(adapted from Forgie and Forgie [6])

2.3 STOP DETERMINATION

If P(i) has not previously been labeled either a fricative or a vowel, it is labeled a stop (i.e., TYPE(i) = 3) if

$A1(i) < 12.$

In other words, A1(i) is required to be in the lowest 20% of all possible A1 amplitudes. The choice of A1 (rather than A2 or A3) seems dictated by the fact that A2 and A3 are normalized with respect to A1 and could exceed the range 0-63; however, A1 is always guaranteed to be in this range.

2.4 NASAL DETERMINATION

If P(i) has not satisfied the tests for either a fricative, vowel, or stop, it is labeled a nasal (i.e., TYPE(i) = 2) if:

(1) $A1(i) \geq 12$

(2) $Z1(i) \leq 5$

(3) $3 \cdot A2(i) \leq A1(i)$

and

(4) $3 \cdot A3(i) \leq A1(i)$.

It is important to recall that a vowel is distinguished by a local maximum. Thus, if P(i) satisfies the amplitude requirements for a vowel but not the duration requirement (i.e., DUR(i) < 8), then a search would be made of neighboring segments for one that is more likely to be a vowel. Such a segment, which could satisfy the tests for both a vowel and a nasal, would always be labeled a vowel. One possible improvement to the system could be made by performing the vowel and nasal tests concurrently rather than serially.

Nakata [7] and Fujimura [8] have experimentally derived characteristic properties of nasals. For example, Figure 5 illustrates a spectral envelope for a typical /m/. Note that the lowest resonant frequency (formant 1) is in the range 200 to 300 Hz, which corresponds to 4-6 in the Vicens-Reddy system.

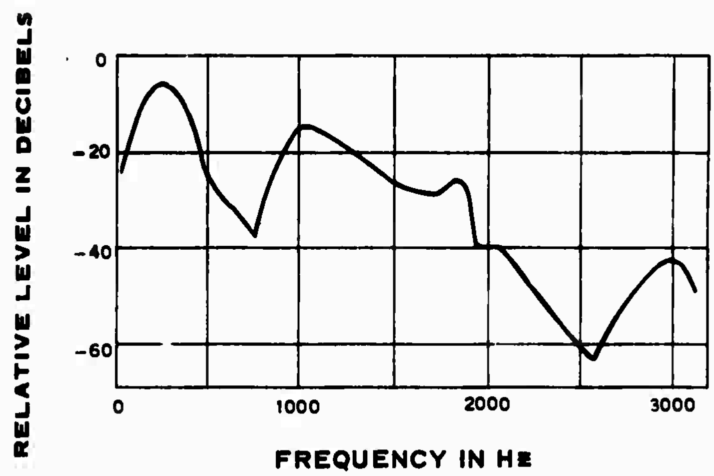


Figure 5. Spectral Envelope for a Typical /m/
(adapted from Nakata [7])

Condition (2) for a nasal requires that $Z1(i) \leq 5$. Since

$$3 \leq Z1(i) \leq 5,$$

we have that Z1(i) is either 3, 4, or 5, which corresponds closely to the range of 4 to 6. The remaining criteria appear to have been developed heuristically and no further explanation will be offered.

It appears from Figure 5 that the process of nasal determination could be improved by adding a requirement that the highest resonant frequency (formant 3) be around 3000 Hz, i.e., that Z3 be approximately 60.

2.5 CONSONANT DETERMINATION

If P(i) has not satisfied the tests for either a fricative, vowel, stop, or nasal, then it is labeled a consonant (i.e., TYPE(i) = 1) if it satisfies the sole condition that it is a sustained segment.

3. SECONDARY CLASSIFICATION

At the completion of primary classification, each P-segment has been labeled. However, the linguistic label "burst" has not yet been assigned. Secondary classification begins by combining appropriate adjacent fricatives and stops. Various fricatives are then identified as "bursts." Next, appropriate transitionals, consonants, nasals, and stops are labeled "burst." Finally, bursts adjacent to other bursts or fricatives may be combined on the basis of tests given below.

The P-matrix is recompacted, and a final determination of the beginning and ending segments is performed.

3.1 COMBINING OF ADJACENT FRICATIVES AND STOPS

Adjacent fricatives and adjacent stops are combined on the basis of the conditions illustrated in Table 2, where it assumed that the operations are performed for i=3, ... , SIZEP.

Table 2. Rules for Combining Adjacent Fricatives and Stops

Condition \ Case	1	2	3	4	5	6	7	8	9	10
TYPE(i)	FRIC	FRIC	FRIC	FRIC	FRIC	STOP	STOP	STOP	STOP	STOP
TYPE(i-1)	FRIC	FRIC	FRIC	FRIC	FRIC	STOP	STOP	STOP	STOP	STOP
CLØ(i)<-12						NO	NO	NO	NO	NO
|DUR(i)-DUR(i-1)|≤4	YES	YES	YES	NO	NO	YES	YES	YES	NO	NO
DUR(i)<DUR(i-1)				YES	NO				YES	NO
NAT(i)	SUST	SUST	TRAN			SUST	SUST	TRAN		
NAT(i-1)	SUST	TRAN				SUST	TRAN			
Actions To Be Performed	1,2,4	1,3,4	1,4	1,4	1,3,4	1,2,4	1,3,4	1,4	1,4	1,3,4

Note: The condition CLØ(i)<-12 appeared in an earlier version of the program as CLØ(i)<-4.

The following actions are to be performed in conjunction with the table:

1. DUR1 = DUR(i) + DUR(i-1).
2. Recompute the parameter values of P(i-1) for A1, Z1, A2, Z2, A3, and Z3. This calculation is shown below, using A1 as an example:

   $A1MN(i-1) = \min\{A1MN(i-1),\ A1MN(i)\},$

   $$A1(i-1) = \frac{A1(i-1)\cdot DUR(i-1) + A1(i)\cdot DUR(i)}{DUR(i-1) + DUR(i)},$$

   $A1MX(i-1) = \max\{A1MX(i-1),\ A1MX(i)\}.$

3. For columns 2 through 22 of the P-matrix, set

   P(i-1) = P(i).
4. DUR(i-1) = DUR1,

   SXT(i-1) = min {SXT(i-1), SXT(i)},

   move all the P-matrix rows up one row, and set

   SIZEP = SIZEP -1.

We shall illustrate the use of this table by considering the following example: suppose that

TYPE(i) = FRIC and TYPE(i-1) = FRIC.

We then check to see if

$|DUR(i) - DUR(i-1)| \le 4$.

If it is, we check NAT(i) and NAT(i-1). If both are SUST, actions 1, 2, and 4 above are performed.

3.2 IDENTIFICATION OF APPROPRIATE FRICATIVES AS BURSTS

For i=2, ... , SIZEP, we label P(i) a burst (i.e., TYPE(i) = 4) if P(i) has previously been labeled a fricative (i.e., TYPE(i) = 5), it satisfies the condition that*

$5 \cdot DUR(i) + 2 \cdot Z3(i) \le 150$,

and either:

(1) $DUR(i) \le 6$,

or

(2) $DUR(i) \ge 5$ and $A3(i) \le A1(i)$,

or

(3) $DUR(i) \ge 5$ and $A3(i) \le A2(i)$.

3.3 IDENTIFICATION OF APPROPRIATE TRANSITIONALS, CONSONANTS, NASALS, AND STOPS AS BURSTS

P(i) (i=2, ... , SIZEP) is labeled a burst (i.e., TYPE(i) = 4) if TYPE(i) = 0, 1, 2, or 3 (i.e., P(i) is already either a transitional, consonant, nasal, or stop) and $Z3(i) \ge 60$ or $A1(i) \le 16$.

---

*In an earlier version of the program, this condition appeared as $5 \cdot DUR(i) + 2 \cdot Z3(i) \le 140$.

However, if either

$Z3(i) < 60$ or $A1(i) > 16$,

and any one of the following six sets of conditions is satisfied, we also label P(i) a burst*:

(1) $40 \le Z3(i) \le 50$,
$Z3(i) + Z2(i) \le 60$,
$A1(i) + A2(i) < 20$, and
$A1(i) \le 6$.

(2) $40 \le Z3(i) \le 50$,
$Z3(i) + Z2(i) \le 60$,
$A1(i) + A2(i) < 20$,
$A1(i) > 6$, and
$A3(i) \ge A1(i)$.

(3) $Z3(i) > 50$,
$A1(i) + A2(i) < 20$, and
$A1(i) \le 6$.

(4) $Z3(i) > 50$,
$A1(i) + A2(i) < 20$,
$A1(i) > 6$, and
$A3(i) \ge A1(i)$.

(5) $Z3(i) \le 50$,
$Z3(i) + Z2(i) > 60$,
$A1(i) + A2(i) < 20$, and
$A1(i) \le 6$.

(6) $Z3(i) \le 50$,
$Z3(i) + Z2(i) > 60$,
$A1(i) + A2(i) < 20$,
$A1(i) > 6$, and
$A3(i) \ge A1(i)$.

---

*The condition $4\emptyset \le Z3(i) \le 50$ in (1) and (2) appeared in an earlier version of the program as $45 \le Z3(i) \le 50$. Also, the condition $A1(i) \le 6$ in (1), (3), and (5) was originally $A1(i) \le 1\emptyset$, and the condition $A1(i) > 6$ in (2), (4), and (6) was originally $A1(i) > 1\emptyset$.

Halle, Hughes, and Radley [9] have noted that stop bursts may be characterized as follows:

/p/ and /b/ (the labial stops) have a high concentration of energy around 500 - 1500 Hz;

/t/ and /d/ (the postdental stops) have either a flat spectrum or have high energy concentrations above 4000 Hz and around 500 Hz;

/k/ and /g/ (the palatal and velar stops) have high concentrations of energy around 1500 - 4000 Hz.

The above data were obtained from an analysis of energy spectra of the phonemes /p/, /b/, /t/, /d/, /k/, and /g/. Closer examination of these spectra reveals that the third formant frequency for /k/ and /g/ is characteristically between 3000 Hz and 4500 Hz, which corresponds to

$$60 \le Z3 \le 90$$

in the Vicens-Reddy system. As stated above, a transitional, consonant, nasal, or stop with the property that

$$Z3 \ge 60$$

is relabeled a burst. In this case, a reasonably close correlation exists. However, no correspondance exists between the remaining tests for a burst and the characterizations given in [9].

3.4 COMBINING BURSTS ADJACENT TO OTHER BURSTS OR FRICATIVES

The entire P-matrix, beginning with P(2) is searched for burst segments. When such a segment has been found, the most adjacent previous segment (which has not been previously combined into another burst segment) is examined to determine whether it is a burst or a fricative. If so, then the burst segment P(i) is combined with the previous segment by adding DUR(i) to the duration of the previous burst or fricative segment. If the previous segment is a burst, then its new duration is tested and, if it is greater than or equal to 80 ms., the TYPE of the segment is changed from 4 (i.e., burst) to 5 (i.e., fricative).

Independent of whether or not P(i) was combined with a previous segment, P(i+1) is examined to determine if it is a burst or a fricative. If it is, then P(i) is combined with P(i+1) by adding DUR(i) to DUR(i+1) and resetting the beginning

Q-segment of P(i+1) to point to the beginning Q-segment of P(i). Again, if P(i+1) is a burst, the new duration is tested and, if it is greater than or equal to 80 ms., TYPE(i+1) is changed from 4 to 5.

A possible result of this combining procedure is that a burst segment P(i) could be combined with a fricative or burst preceding it and also with a fricative or burst following it. The resulting duration of the two segments would then be erroneous. A more detailed description of this procedure can be found in Figure 6.

3.5 DETERMINATION OF BEGINNING AND ENDING SEGMENTS

The P-matrix is recompacted by suppressing all segments P(i) for which TYPE(i) = -1 (recall that all segments so flagged were previously combined with adjacent segments). Let k denote the row number of the last row of the recompacted P-matrix.

Beginning with P(k), the P-matrix is examined backwards from i = k to i = 2 as follows: if either:

(1) P(i) is a stop

or

(2) P(i) is not a stop, burst, or fricative but

$$4 \cdot DUR(i) + 2 \cdot A1(i) < 36,$$

then P(i-1) is examined similarly until we find a P(i) which satisfies neither (1) nor (2). Such a P-segment is either:

(1) a burst or

(2) a fricative or

(3) not a stop, burst, or fricative but

$$4 \cdot DUR(i) + 2 \cdot A1(i) \geq 36.$$

If P(i) is a burst, then we set

$$SIZEP = i \text{ and } DUR(i) = \max\{6, \tfrac{1}{2} DUR(i)\}.$$

This implies that the ending segment is a burst of duration not less than 60 ms.

If P(i) is a fricative, then we set

$$SIZEP = i \text{ and } DUR(i) = \min\{12, DUR(i)\}.$$

This implies that the ending segment is a fricative of duration less than or equal to 120 ms. If DUR(i) is now ≤ 10

or

$$5 \cdot DUR(i) + Z3(i) \leq 110$$

or

$$A1(i) + A2(i) < 8,$$

then we set

$$TYPE(i) = 4,$$

i.e., the fricative P(i) is relabeled a burst.

If P(i) is not a stop, burst, or fricative but

$$4 \cdot DUR(i) + 2 \cdot A1(i) \geq 36,$$

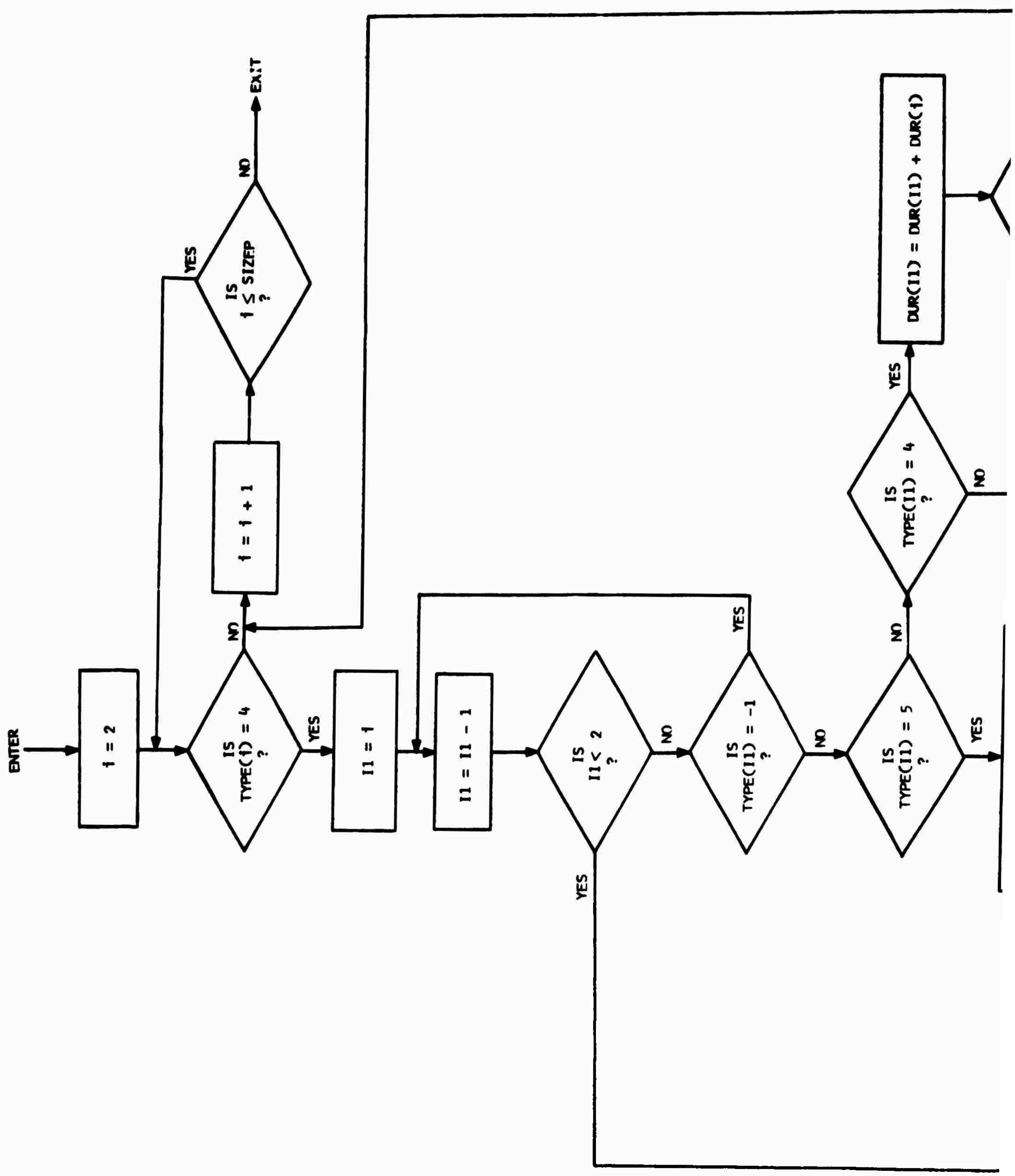
then we set

$$SIZEP = \min\{k, i+1\}.$$

This means that if $i = k$, the speech sample ends with a consonant, nasal, vowel, or transitional. However if $i < k$, the sample ends with the segment following P(i). This may be a vowel, nasal, or consonant which did not pass the test, or a stop.

To determine the beginning segment of the P-matrix, if P(2) is a stop and DUR(2) is greater than 5, then the beginning Q-segment of P(2) is defined to be SBG(2) = SBG(2) + DUR(2) -5 and we set DUR(2) = 5.

A more detailed description of the handling of beginning and ending segments can be found in Figure 7.


ENTER
i = 2
IS
TYPE(i) = 4
?
YES
NO
i = i + 1
IS
i ≤ SIZEP
?
YES
NO
EXIT
I1 = i
I1 = I1 - 1
IS
I1 < 2
?
YES
NO
IS
TYPE(I1) = -1
?
YES
NO
IS
TYPE(I1) = 5
?
YES
NO
IS
TYPE(I1) = 4
?
YES
NO
DUR(I1) = DUR(I1) + DUR(i)

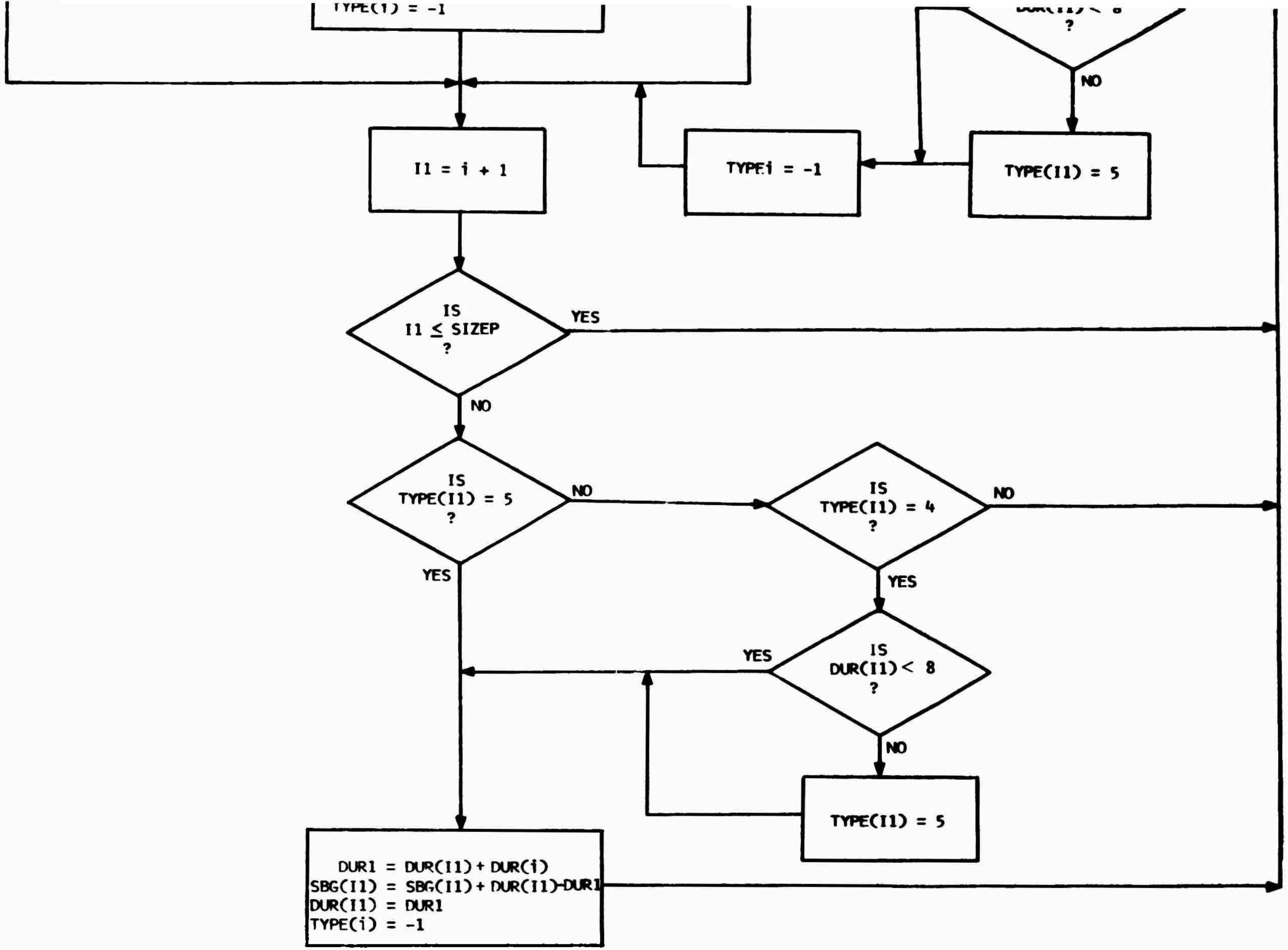
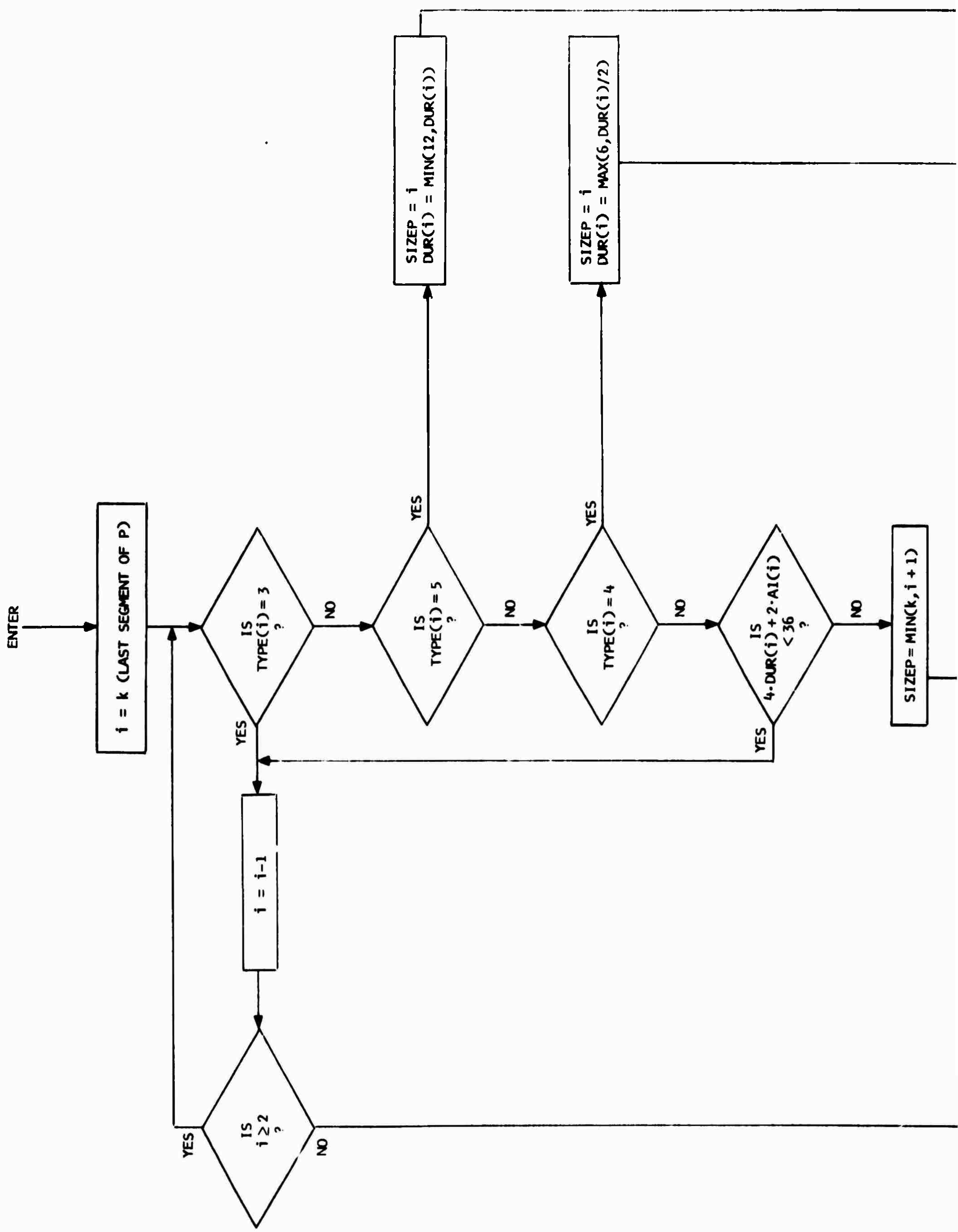


Figure 6. **Flow Chart for Combining Bursts Adjacent to Other Bursts or Fricatives**


ENTER
i = k (LAST SEGMENT OF P)
IS TYPE(i) = 3 ?
YES
NO
IS TYPE(i) = 5 ?
YES
NO
SIZEP = i
DUR(i) = MIN(12,DUR(i))
IS TYPE(i) = 4 ?
YES
NO
SIZEP = i
DUR(i) = MAX(6,DUR(i)/2)
IS 4·DUR(i) + 2·AI(i) < 36 ?
YES
NO
SIZEP = MIN(k, i + 1)
i = i-1
IS i ≥ 2 ?
YES
NO

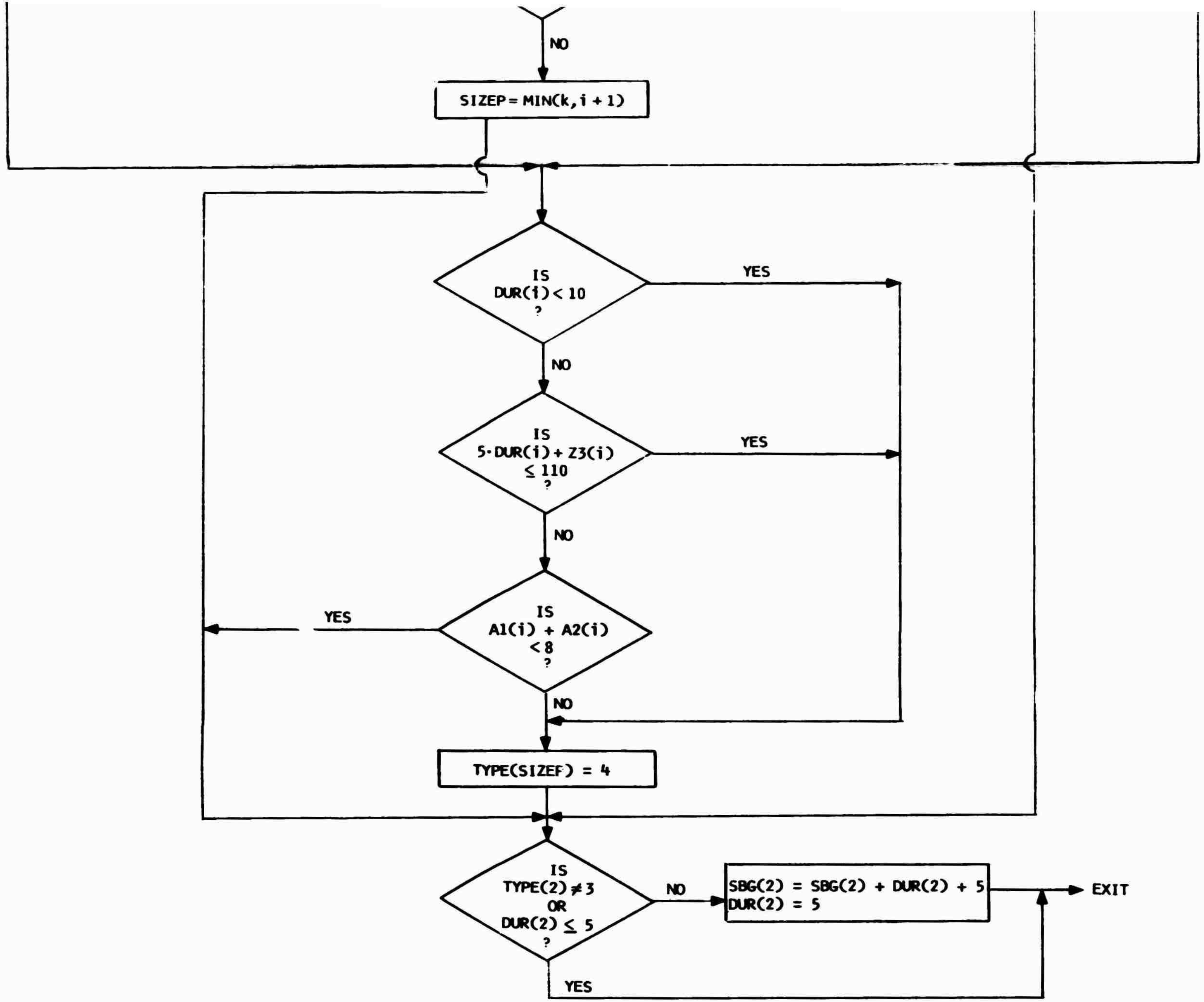


Figure 7. Flow Chart for Determination of Beginning and Ending Segments

4. CONSTRUCTION OF THE R-MATRIX

The results of primary and secondary classification will now be used to construct an array called the R-matrix, or feature matrix. This matrix will be used in the lexicon lookup portion of the program to identify a spoken message.

The R-matrix consists of 10 columns and a maximum of 40 rows. Let $R = (r_{ij})$, where $i = 1, \ldots, m$ and $j = 1, \ldots, 10$, where $m \leq 40$. The first row of R is defined as follows:

$r_{1,1}$ = number of vowels in the message,

$r_{1,2}$ = number of fricatives in the message,

$r_{1,3}$ = an unused position of the array,

$r_{1,4}$ = m +1,

$r_{1,5}$ = row number of first* vowel appearing in message,

$r_{1,6}$ = row number of second vowel appearing in message,

$r_{1,7}$ = row number of third vowel appearing in message,

$r_{1,8}$ = row number of fourth vowel appearing in message,

$r_{1,9}$ = row number of fifth vowel appearing in message,

$r_{1,10}$ = an octal pattern representing the sequence of vowels and fricatives in the message; an octal "1" represents a vowel, and an octal "2" represents a fricative.

---

*If the message contains only one vowel, $r_{1,6} = r_{1,7} = r_{1,8} = r_{1,9} = 0$.

The remaining rows of the R-matrix are defined as follows for $i = 2, \ldots, m$:

$r_{i,1}$ = alphanumeric phonemic label of P(i) (see Table 1 for the four-character phonemic labels),

$r_{i,2}$ = DUR(i), the length of P(i) in minimal segments,

$r_{i,3}$ = A1(i),

$r_{i,4}$ = Z1(i),

$r_{i,5}$ = A2(i),

$r_{i,6}$ = Z2(i),

$r_{i,7}$ = A3(i),

$r_{i,8}$ = Z3(i),

$r_{i,9}$ = SXT(i).

APPENDIX

Vowel Phonemes as Adapted from Reddy [10]

PHONEME	AS IN
i	eve
I	it
ɛ	met
æ	at
ɝ	bird
ʌ	up
a	father
ɔ	all
u	foot
ʊ	boot

Note: e as in "mate" and o as in "obey" are not included because they are considered to be diphthongs.

REFERENCES

1. P. Vicens, Aspects of Speech Recognition by Computer, Stanford University, Memo AI-85 (CS 127), 1969, 210 pp.

2. I. Kameny and H. B. Ritea, Description and Analysis of the Vicens-Reddy Preprocessing and Segmentation Algorithms, System Development Corporation, Technical Memorandum TM-4652/200/00, 4 December 1970, 63 pp.

3. G. W. Hughes and M. Halle, Spectral Properties of Fricative Consonants, J. Acoust. Soc. Amer., 28(1956), 303-310.

4. J. M. Heinz and K. N. Stevens, On the Properties of Voiceless Fricative Consonants, J. Acoust. Soc. Amer., 33(1961), 589-596.

5. G. E. Peterson and H. L. Barney, Control Methods Used in a Study of the Vowels, J. Acoust. Soc. Amer., 24(1952), 175-184.

6. J. W. Forgie and C. D. Forgie, Results Obtained from a Vowel Recognition Computer Program, J. Acoust. Soc. Amer., 31(1959), 1480-1489.

7. K. Nakata, Synthesis and Perception of Nasal Consonants, J. Acoust. Soc. Amer., 31(1959), 661-666.

8. O. Fujimura, Analysis of Nasal Consonants, J. Acoust. Soc. Amer., 34(1962), 1865-1875.

9. M. Halle, G. W. Hughes, and J.-P. A. Radley, Acoustic Properties of Stop Consonants, J. Acoust. Soc. Amer., 29(1957), 107-116.

10. D. R. Reddy, An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave, Stanford University, Memo AI-43 (CS 49), 1966, 144 pp.

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) System Development Corporation Santa Monica, California	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP
3. REPORT TITLE Description and Analysis of the Vicens-Reddy Recognition Algorithms	
4. DESCRIPTIVE NOTES (Type of report and inclusive dates) Technical Report -- Nov 1970 - March 1971	
5. AUTHOR(S) (First name, middle initial, last name) Iris Kameny H. Barry Ritea	

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
29 March 1971	29	10

8a. CONTRACT OR GRANT NO. DAHC15-67-C-0149 b. PROJECT NO. ARPA Order #1327, Amendment #3, Program Code No. 1D30, and 1P10. c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) TM-4652/300/00 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
10. DISTRIBUTION STATEMENT Distribution of this document is unlimited.	
11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY

13. ABSTRACT

This document provides a detailed description and analysis of the recognition algorithms used in the Vicens-Reddy speech recognition system.

DD FORM 1 NOV 65 1473

UNCLASSIFIED

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Automatic Speech Recognition						
Acoustic Phonetics						